

Non-standard texts: from theoretical positions to Natural Language Processing normalisation

Cédric Lopez* Mathieu Roche** Rachel Panckhurst***



*R&D, Viseo, Grenoble cedric.lopez@viseo.com
**UMR TETIS, Cirad, Irstea, AgroParisTech, Montpellier; LIRMM, UMR 5506 CNRS & Université Montpellier mathieu.roche@cirad.fr
***Praxiling UMR 5267 CNRS & Université Paul-Valéry Montpellier 3 rachel.panckhurst@univ-montp3.fr

theoretical position

- › annotation is not neutral
- › annotation is linked to interpretative frameworks
- › researchers should not be trapped
- › researchers need to conduct own annotation
- › full corpus (88,000 sms) & samples available for download

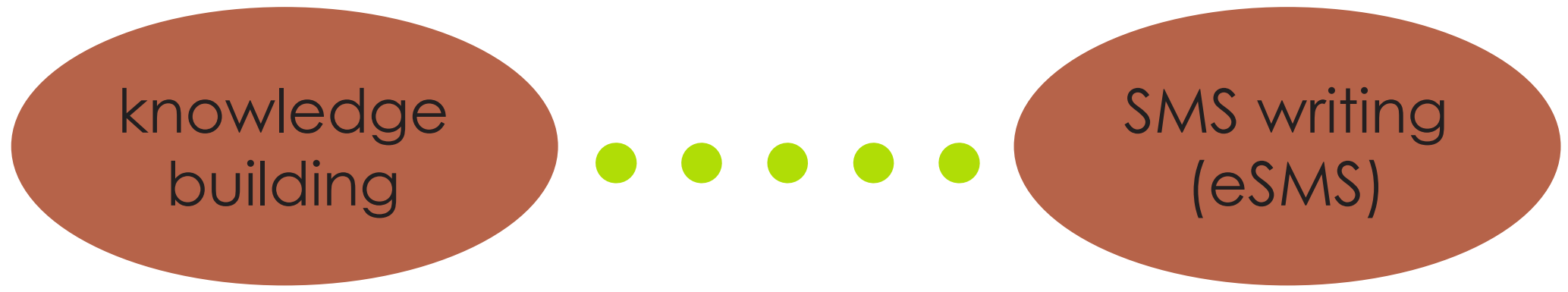
88milSMS corpus

alignment algorithms for transcoding

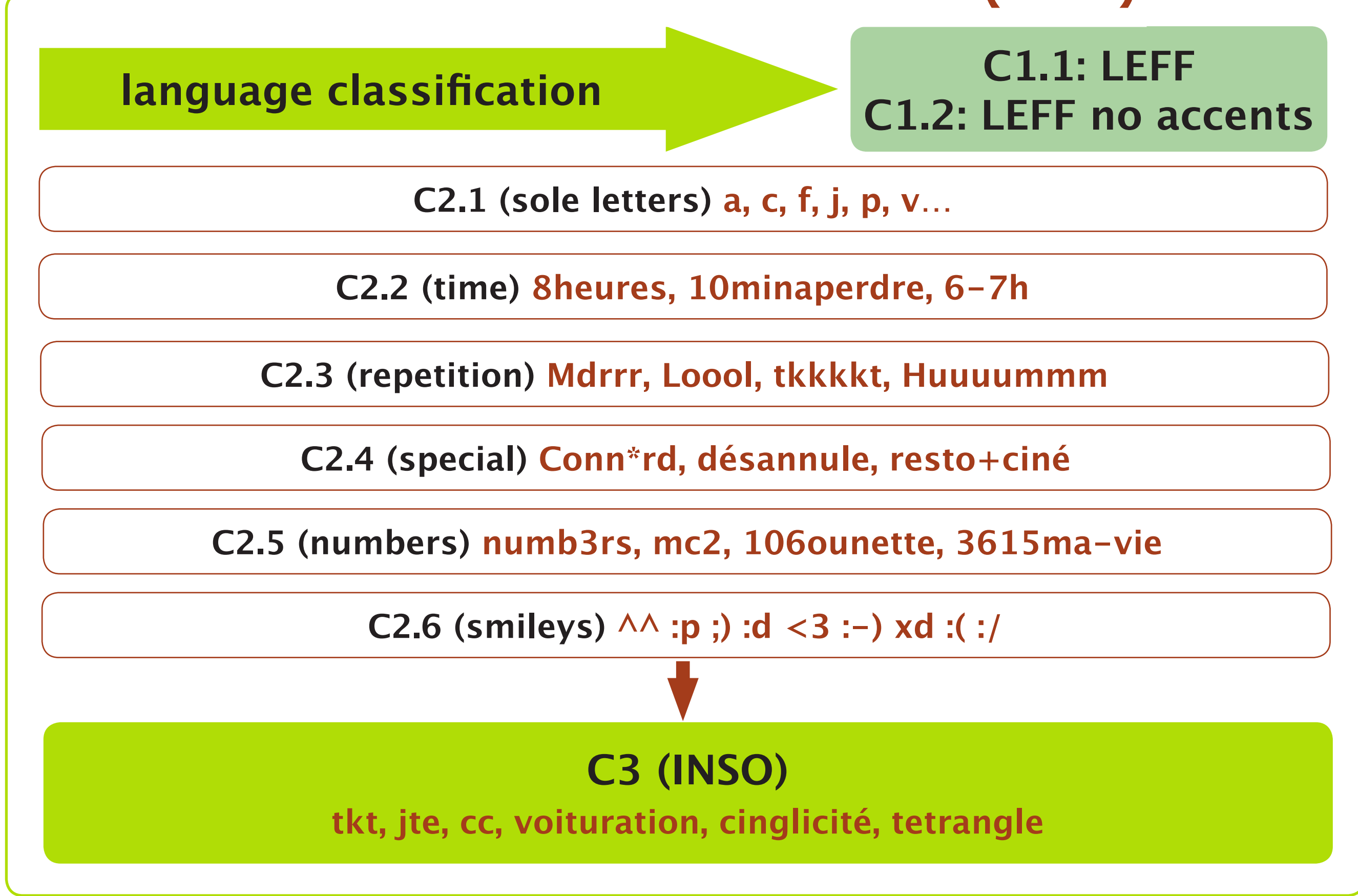
Case 1		
plus	svt	possible
plus	souvent	possible

Case 2		
Vasi	lâche	moi
Vas-y	lâche-moi	

Case 3		
T'as		eu
Tu	as	eu



'unknown' non-standard items (INSO)



automatic normalisation techniques

- › new typology of detected 'mistakes'
- › normalisation based on most frequent errors
 - › confrontation with:
 - traditional automatic translation,
 - speech recognition,
 - spelling/grammatical checker principles
- › comparison between different types of instant media (SMS, forums, tweets)

PLIN Day, Louvain-la-Neuve, 12 May 2016